Research Report

Applying Multiphase Sampling to Selecting Testlets With Constraints on Item Blocks

ETS RR–19-03

Jiahe Qian
Lixiong Gu
Shuhong Li

December 2019

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Applying Multiphase Sampling to Selecting Testlets With Constraints on Item Blocks

Jiahe Qian, Lixiong Gu, & Shuhong Li

Educational Testing Service, Princeton, NJ

In assembling testlets (i.e., test forms) with a pool of new and used item blocks, test security is one of the main issues of concern. Strict constraints are often imposed on repeated usage of the same item blocks. Nevertheless, for an assessment administering multiple testlets, a goal is to select as large a sample of testlets as possible. In this study, the algorithm of multiphase sampling was applied to selecting and augmenting the sample of testlets to be administered. Several topics related to the algorithm are discussed, such as the termination of the algorithm, the dynamics of sample size, and the effectiveness of the algorithm. A real database of testlets was used in the study.

Keywords Testlet; multiphase sampling; constraints on testlets; sample augmentation; termination of the algorithm

doi:10.1002/ets2.12239

Item blocks are often used to assemble test forms in large-scale assessments, such as in state accountability tests and international English proficiency assessments. Each block abides by specific content and statistical specifications, and multiple sets of item blocks can be constructed and used interchangeably in the test. Such a block-based design provides the means for equating and monitoring trends; moreover, it improves efficiency and is cost efficient through the reuse of item blocks. In this study, *testlet* refers to a test form that consists of several item blocks; the blocks can be passage based, and each passage comprises a set of items. A testlet is defined as a group of items related to a single content area that is developed as a unit and contains a fixed number of predetermined paths that a test taker may follow (Wainer & Kiely, 1987; Wainer & Lewis, 1990). The items in all the blocks of a testlet are administered to test takers together.

The criterion for a reasonable assembling process is providing normed test forms (testlets) that are parallel and exchangeable in psychometric properties (Lord, 1980) to test takers; that is, the assembled test forms are equivalent in terms of their statistical and content-related properties (Chen, Chang, & Wu, 2012). The statistical properties can be based on item response theory (IRT), such as the deviation of average *a* parameters of blocks; the test information function (TIF; van der Linden, 2005); and the test characteristic curve (TCC; Belov & Armstrong, 2008). Based on the TIF and TCC, combinatorial optimal methods can also be used in attaining parallel test forms (Ali & van Rijn, 2016; Debeer, Ali, & van Rijn, 2017).

In addition to optimization, another concern for test assembly is test security (Bennett, 1998; Foster & Miller, 2012; Wollack & Fremer, 2013). Test security is likely to be compromised if item blocks are overly reused. If the blocks with good psychometric properties are overused, the exposure rate can be high (Chang & Zhang, 2002; Stocking & Lewis, 1998), exacerbating the likelihood of hurting test security (Luecht, 2012). Underused or never-used blocks lead to inefficiency and should be avoided. Therefore some constraints must be imposed on the reused blocks in the test assembly or sampling process, as proposed in this study. Compared with the constraints imposed on computer-based testing (Chang, Qian, & Ying, 2001; van der Linden, 2003; Veldkamp & van der Linden, 2002), the constraints set for the blocks are relatively more straightforward.

When assembling forms for an assessment in which frequent tests are administered, it is desirable to acquire a large enough testlet database so that there can be sufficient forms to assess the appropriateness of the content components and to have enough backup test forms for substitution. However, this is a demanding task, because the constraints are primarily imposed on the blocks, whereas the unit selected from the database is a whole test form or testlet. Moreover, the distributions of the blocks across the forms can be quite uneven. In the database used in this study, one block for one

Corresponding author: J. Qian, E-mail: jqian@ets.org

position appears in more than 6,000 testlets, whereas four other blocks for the same position appear in only 1 testlet (see details in the section titled "Database and Symbol Definitions").

For a database with such imbalance, a regular sampling approach, such as simple random sampling (Cochran, 1977; Kish, 1965) of testlets, is obviously not a good strategy, because the imbalance will cause inefficiency in sampling. When simple random sampling is used in selecting testlets from the database, some blocks may be oversampled, while others are undersampled. To comply with the imposed constraints, most of the testlets drawn in a sample need to be stripped because of the replica in blocks across testlets; for the details of stripping, see the section titled "Process of Stripping the Replica Blocks." Thus a refined sampling strategy needs to be developed, and a stripping process is integrated into the algorithm proposed in this study to produce test forms that conform to all imposed constraints.

The goal of this study was to propose a multiphase sampling process based on iterative algorithms in selecting testlets to assemble as large a sample as possible of testlets that meet the constraints imposed on the blocks. The effectiveness of the algorithm of the assembly engine is proved when the algorithm ends its selection process; moreover, there will be no testlets that meet the constraints imposed on blocks remaining in the database. Several primary properties of the algorithm are also provided, including the criterion for termination of the algorithm and the dynamics in sample size. All the testlets drawn by the algorithm in the sample are approximately equivalent and meet specifications. Such a sample represents a subdatabase of testlets ready to be administered. A real database of test forms was used for the study.

There is almost no research on assembling testlets by applying sampling techniques for simultaneously creating a sample of all possible testlets with constraints. In the sample of selected testlets, there were no issues of overusing or underusing blocks. In addition, there were no unused blocks left in the database, so the testing resources were fully utilized. The algorithm also assembles as large a sample of testlets as possible, while other computational methods for automated test assembly usually create one test form in one run. This kind of computational method can be based on the technique of mixed-integer programming (Luecht & Hirsch, 1991, 1992; van der Linden, 1998) or the weighted deviation model (Swanson & Stocking, 1993) using a heuristic approach to identify acceptable sets of items for a test.

Although optimization is not the focus of this report, the algorithm can be improved for optimization under constraints. Optimal tactics, such as the minimum deviation of average *a* parameters of blocks, can be readily embedded in the multiphase sampling process, and a Bellman equation (Bellman, 1957) can be employed to achieve optimization for the blocks across positions. When the TIF and the TCC of blocks are available, the methods of combinatorial optimization can also be used in attaining parallel test forms. Furthermore, the method can be modified to be used with other types of testlets, though the algorithm was developed for the block-based testlets.

In the next section, the database of block-based testlets and the blocks on the testlets are introduced, including the symbols used to define various concepts and the constraints imposed on the sampling process. In the "Method" section, the methodologies applied are reviewed and evaluated, including the partitioning of the block sets, the creation of an initial sample, and the algorithm of multiphase sampling. The properties of the algorithm, such as the termination of the algorithm, the dynamics in sample size, and the effectiveness of the algorithm, are also discussed. In the "Results" section, some empirical results are presented. The final section contains a summary and conclusion.

Database of Block-Based Testlets

Database and Symbol Definitions

To describe the general problem under study, the following example of a large-scale language assessment is considered. The units of the database are testlets; each testlet consists of three distinct blocks of 14 items in successive position order, called positions R1, R2, and R3, respectively.

Let $\mathbb{B}1$, $\mathbb{B}2$, and $\mathbb{B}3$ be the sets of blocks for Positions 1, 2, and 3, respectively. Let $\mathbb{b}_\alpha$, $\mathbb{b}_\beta$, and $\mathbb{b}_\gamma$ be the symbols of the blocks in $\mathbb{B}1$, $\mathbb{B}2$, and $\mathbb{B}3$. Let $\mathbb{f}_{\alpha\beta\gamma} = \left(\mathbb{b}_\alpha|\mathbb{b}_\beta|\mathbb{b}_\gamma\right)$ be a testlet with blocks $\mathbb{b}_\alpha \in \mathbb{B}1$, $\mathbb{b}_\beta \in \mathbb{B}2$, and $\mathbb{b}_\gamma \in \mathbb{B}3$, respectively. Let $\mathfrak{A}_1$, $\mathfrak{A}_2$, and $\mathfrak{A}_3$ be the index sets of the blocks in $\mathbb{B}1$, $\mathbb{B}2$, and $\mathbb{B}3$, respectively. Thus $\mathbb{B}1 = \left\{\mathbb{b}_\alpha|\alpha \in \mathfrak{A}_1\right\}$, $\mathbb{B}2 = \left\{\mathbb{b}_\beta|\beta \in \mathfrak{A}_2\right\}$, and $\mathbb{B}3 = \left\{\mathbb{b}_\gamma|\gamma \in \mathfrak{A}_3\right\}$.

Figure 1 shows three sets of blocks in three positions: Positions 1 – 3 contained in the example used in this study. For each position, a number of blocks are gathered, and each block meets the same content and statistical requirements and thus can be used interchangeably. Because of the concerns that test speededness might affect item parameters, the blocks

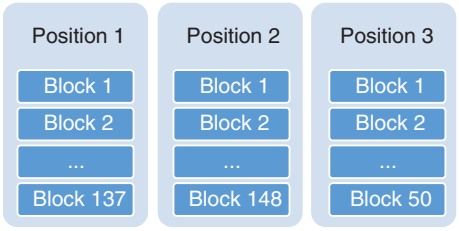

Figure 1 Three sets of item blocks in Positions 1 – 3.

Table 1 Distribution of Blocks Used in Position 1 and Block Average Item Parameters in Testlet Database $\mathbb{U}$

Index α of block $\mathbb{b}_\alpha$	Frequency	Mean_a[a]	Mean_b[a]	O index[b]
1	5	0.7868	−1.0364	c
2	2,555	0.6166	−0.9503	c
3	1,150	0.6625	−0.8775	
4	1,322	0.7376	−0.9605	
5	468	0.7685	−0.7379	c
6	211	0.7945	−1.0403	
7	387	0.8170	−0.8723	
8	2,244	0.6934	−1.4365	c
9	2,154	0.6699	−0.8755	c
10	907	0.6866	−0.8906	c
...	...	...	...	...
133	3	0.7686	−0.9363	c
134	7	0.7343	−1.0892	c
135	2	0.6988	−0.7385	c
136	5	0.7775	−1.0523	c
137	118	0.7109	−0.8423	c
Average		0.6750	−0.7398	

[a]Mean of the estimated a or b parameters of the items in each block. Note that only 6 item blocks (out of 137) are unique in $\mathbb{B}1$. The rest (i.e., 131 blocks) are common with $\mathbb{B}2$ blocks. [b]Index for overlapping. [c]Block overlaps with one of the blocks in $\mathbb{B}2$.

in $\mathbb{B}3$ are immovable; that is, the blocks in $\mathbb{B}3$ cannot exchange positions with those in $\mathbb{B}1$ or $\mathbb{B}2$. The blocks in $\mathbb{B}1$ and $\mathbb{B}2$ are allowed to exchange positions with each other.

In Figure 1, there are 137 blocks in $\mathbb{B}1$ for Position 1, as shown in Table 1; for Positions 2 and 3, there are 148 and 50 blocks in $\mathbb{B}2$ and $\mathbb{B}3$, provided in Tables 2 and 3, respectively. There are 131 overlapping blocks across $\mathbb{B}1$ and $\mathbb{B}2$, which means they are available to be selected in either position ($\mathbb{B}1$ or $\mathbb{B}2$). It is worth mentioning that four blocks in $\mathbb{B}3$ (i.e., $\gamma = 1$, $\gamma = 26$, $\gamma = 29$, and $\gamma = 47$) are only used once in a single testlet (see the distribution of the 50 blocks in $\mathbb{B}3$ in Table 3). Obviously, the database is characterized with uneven distributions of blocks across forms.

Let $\mathbb{U} = \left\{\mathbb{f}_{\alpha\beta\gamma}\right\}$ be the database of all testlets. Thus

$$\mathbb{U} = \left\{\left(\mathbb{b}_\alpha|\mathbb{b}_\beta|\mathbb{b}_\gamma\right)|\alpha \in \mathfrak{Q}_1, \beta \in \mathfrak{Q}_2, \gamma \in \mathfrak{Q}_3\right\}. \quad (1)$$

Because the numbers of blocks for $\mathbb{B}1$, $\mathbb{B}2$, and $\mathbb{B}3$ are 137, 148, and 50, respectively, the total number of all possible testlets in $\mathbb{U}$ equals 1,013,800 (= 137 · 148 · 50) if no constraints are imposed.

In developing assessments with multiple forms, some blocks are not compatible for use in a single testlet, and certain constraints are therefore imposed (see the next section). The testlets in $\mathbb{U}$ are to be screened specifically so that there are no undesirable properties among them (Wainer & Lewis, 1990). Define $\mathbb{F}$ as a subset of $\mathbb{U}$, $\mathbb{F} \subseteq \mathbb{U}$, that contains all the eligible forms that can be used in this study. All the testlets in $\mathbb{F}$ are filtered by a set of content and psychometric rules (Mislevy, 2006) and are ready to be administered operationally. The total number of forms in $\mathbb{F}$ is 81,731, which accounts for 8.1% of the size of $\mathbb{U}$. The rest of the testlets were filtered out from $\mathbb{U}$ by the content or psychometric requirements. For example, testlets with overlapping or conflicting content across blocks or items (referred to as enemy items/blocks) are all excluded (Chen et al., 2012).

Table 2 Distribution of Blocks Used in Position 2 and Block Average Item Parameters in Testlet Database $\mathbb{U}$

Index β of block $\mathbb{b}_\beta$	Frequency	Mean_*a*[a]	Mean_*b*[a]	O index[b]
1	213	0.7868	−1.0364	c
2	181	0.5759	−0.5147	
3	160	0.6228	−0.4694	
4	60	0.8885	−0.8889	
5	45	0.6166	−0.9503	c
6	29	0.6554	−1.0985	
7	13	0.6405	−0.6465	
8	14	0.8246	−0.7643	
9	65	0.7685	−0.7379	c
10	217	0.6562	−0.9477	
...	...	...	...	...
144	57	0.7974	−0.7849	
145	156	0.6988	−0.7385	c
146	289	0.7775	−1.0523	c
147	85	0.6476	−0.7001	
148	569	0.7109	−0.8423	c
Average		0.6738	−0.7333	

[a]Mean of the estimated *a* or *b* parameters of the items in each block. Note that only 6 item blocks (out of 137) are unique in $\mathbb{B}1$. The rest (i.e., 131 blocks) are common with $\mathbb{B}2$ blocks. [b]Index for overlapping. [c]Block overlaps with one of the blocks in $\mathbb{B}2$.

Table 3 Distribution of Blocks Used in Position 3 and Block Average Item Parameters in Testlet Database $\mathbb{U}$

Index γ of block $\mathbb{b}_\gamma$	Frequency	Mean_*a*[a]	Mean_*b*[a]
1	1	0.5462	−0.6199
2	3,624	0.5955	−0.4612
3	2,481	0.6249	−0.4921
4	840	0.5503	−0.5130
5	2,711	0.7125	−0.5383
6	2,512	0.6879	−0.4832
7	3,594	0.6836	−0.7709
8	3,239	0.7535	−0.7400
9	2,621	0.5685	−0.3526
10	1,003	0.6099	−0.4908
...	...	...	...
46	2,391	0.5921	−0.5631
47	1	0.5399	−0.2262
48	204	0.5976	−0.3029
49	478	0.7018	−0.7209
50	461	0.6506	−0.6015
Average		0.6454	−0.5058

[a]Mean of the estimated *a* or *b* parameters of the items in each block.

Constraints Imposed on Testlets

In this study, two types of constraints were imposed on the blocks in selecting testlets. First, constraints were imposed to avoid blocks being overused or underused and in consideration of test security (Davis & Dodd, 2003; Leung, Chang, & Hau, 2002; van der Linden & Veldkamp, 2004). Second, constraints were proposed for optimization, such as those constraints imposed on the item IRT parameters across blocks. This is discussed in the "Creation of an Initial Sample" section.

For test security, constraints are used to restrict the replica of blocks in each position:

1 Each block in the R1 position can appear only once in the sample of selected forms. No replica of the same $\mathbb{B}1$ block is permitted. The same constraints are imposed on the blocks in $\mathbb{B}2$.

2 Because there are 131 blocks overlapping across $\mathbb{B}1$ and $\mathbb{B}2$, such a block can be used in different testlets, but not in the same one. Thus, for any block in $\mathbb{B}1$ and $\mathbb{B}2$, the total number of its appearances in all the sampled testlets is no more than two.

3 The blocks in $\mathbb{B}3$ can only appear in the R3 position, yet each block is allowed to appear twice. Because among the 50 blocks in $\mathbb{B}3$, 4 blocks appear in one and only one testlet, the maximum size of a batch of sampled testlets is 96 rather than 100. Otherwise, by the pigeonhole principle (Rittaud & Heeffer, 2014), the fact that four $\mathbb{B}3$ blocks appear in one and only one testlet would be contradicted.

Index Set of Sampled Blocks

To control the selection process of multiple phase sampling, each block set—either $\mathbb{B}1$, $\mathbb{B}2$, or $\mathbb{B}3$—is partitioned based on the status of the blocks being used in the testlets already sampled in prior phases. These partitions are denoted with index sets. For example, in $\mathbb{B}1$, let $\mathfrak{q}_{1,1} \subseteq \mathfrak{Q}_1$ be the index set of the blocks that are sampled only once.

Similarly, in $\mathbb{B}2$ and $\mathbb{B}3$, define $\mathfrak{q}_{2,1} \subseteq \mathfrak{Q}_2$ and $\mathfrak{q}_{3,1} \subseteq \mathfrak{Q}_3$. Because the constraints imposed on the blocks in $\mathbb{B}3$ allow each block to be used up to two times, $\mathfrak{q}_{3,2} \subseteq \mathfrak{Q}_3$ is defined as the index set of the blocks that are used exactly twice in testlets. Analogously, we can define sets $\mathfrak{q}_{1,k}$, $\mathfrak{q}_{2,k}$, and $\mathfrak{q}_{3,k+1}$, where k is a natural number that is larger than 1; however, due to the constraints imposed in this study, the sets of $\mathfrak{q}_{1,k}$, $\mathfrak{q}_{2,k}$, and $\mathfrak{q}_{3,k+1}$ $(\forall k > 1)$ are all empty. The main purpose of introducing the index sets is to create the source sets of testlets employed in multiphase sampling.

Method

Creating an Initial Sample

In this study, the initial sample $\mathbb{S}^{0*}$ consists of two subsamples $\mathbb{S}_1^*$ and $\mathbb{S}_2^*$ of testlets. First, select two testlets for each $\mathbb{B}3$ block and create the subsample $\mathbb{S}_1^*$, that is, a set of 96 testlets from the database used in this study. Second, select one testlet for each $\mathbb{B}1$ block and create the subsample $\mathbb{S}_2^*$, a set of 137 testlets from the database. Then, create the union of two sets $\mathbb{S}_1^*$ and $\mathbb{S}_2^*$: $\mathbb{S}^{0*} = \mathbb{S}_1^* \cup \mathbb{S}_2^*$, with a sample size of 233.

In selecting $\mathbb{S}_1^*$ and $\mathbb{S}_2^*$, an optimal strategy is employed to achieve a balance in the IRT parameters for the items in selected blocks. In this study, the database contains the mean of the estimated a or b parameters of the items for each of its blocks; see the mean_a and mean_b columns in Tables 1–3. They are defined as $a_{\mathbb{b}_\alpha}$, $a_{\mathbb{b}_\beta}$, and $a_{\mathbb{b}_\gamma}$ for the blocks $\mathbb{b}_\alpha$, $\mathbb{b}_\beta$, and $\mathbb{b}_\gamma$ in $\mathbb{B}1$, $\mathbb{B}2$, and $\mathbb{B}3$, respectively.

The overall averages of $a_{\mathbb{b}_\alpha}$, $a_{\mathbb{b}_\beta}$, and $a_{\mathbb{b}_\gamma}$ across blocks are $\overline{a}_{\mathbb{b}_{\alpha\cdot}}$, $\overline{a}_{\mathbb{b}_{\alpha\cdot}}$, and $\overline{a}_{\mathbb{b}_{\gamma\cdot}}$, respectively; for example, $\overline{a}_{\mathbb{b}_{\alpha\cdot}}$ is 0.6756, according to Table 1. Deviation of a block can be defined as the difference between its mean-a parameter and the overall average, such as $d_{\mathbb{b}_\alpha} = a_{\mathbb{b}_\alpha} - \overline{a}_{\mathbb{b}_{\alpha\cdot}}$ in $\mathbb{B}1$. The same definitions are used for the blocks in $\mathbb{B}2$ and $\mathbb{B}3$. Given a block $\mathbb{b}_{\alpha_0} \in \mathbb{B}1$, $\left\{ \left(\mathbb{b}_{\alpha_0} | \mathbb{b}_\beta | \mathbb{b}_\gamma \right) \in \mathbb{F} \right\}$ is the set of testlets that contained $\mathbb{b}_{\alpha_0}$. Thus we choose a testlet from the set of $\left\{ \left(\mathbb{b}_{\alpha_0} | \mathbb{b}_\beta | \mathbb{b}_\gamma \right) \in \mathbb{F} \right\}$ satisfying certain conditions on deviations for $\mathbb{B}2$ and/or $\mathbb{B}3$ blocks. One optimal tactic is to select a testlet from the set with block $\mathbb{b}_\beta$ that has the *minimum deviation*; this tactic allowed the testlets in the sample to have comparable median psychometric properties in the IRT a parameters. If more information, such as timing data, is available, more factors can be included to achieve improved optimization in assembling. When the deviations, such as $d_{\mathbb{b}_\alpha}$ and $d_{\mathbb{b}_\beta}$, are available, to obtain balanced properties for blocks across positions, an alternative to the selection strategy can also be considered. For a given block $\mathbb{b}_{\alpha_0} \in \mathbb{B}1$, the altered strategy is to choose a testlet from the set of $\left\{ \left(\mathbb{b}_{\alpha_0} | \mathbb{b}_\beta | \mathbb{b}_\gamma \right) \in \mathbb{F} \right\}$ satisfying the constraints expressed in the Bellman equation on IRT item parameters:

$$\arg \min_{\forall \beta \notin \mathfrak{q}_{2,1}^{t-1}} \left(d_{\mathbb{b}_{\alpha_0}} + d_{\mathbb{b}_\beta} \right), \tag{2}$$

where $t-1$ indicates the cycle of the sampling phase. The solution to Equation 2 can be obtained from the finite testlets in $\mathfrak{q}_{2,1}^{t-1}$; no complex computations, such as mixed-integer linear programming, are needed in obtaining the solution. For example, if the deviation $d_{\mathbb{b}_{\alpha_0}}$ is assumed to be positive, the strategy can choose a testlet with negative $d_{\mathbb{b}_\beta}$ if such a testlet is available. For IRT-based optimal tactics, the TIF or TCC can be used as the statistical target in assembling parallel test forms. Because most of these assembly designs use items or blocks to assemble test forms, such optimal test assembly is constrained to a mixed-integer programming problem (Luecht & Hirsch, 1991). The problem can be addressed using

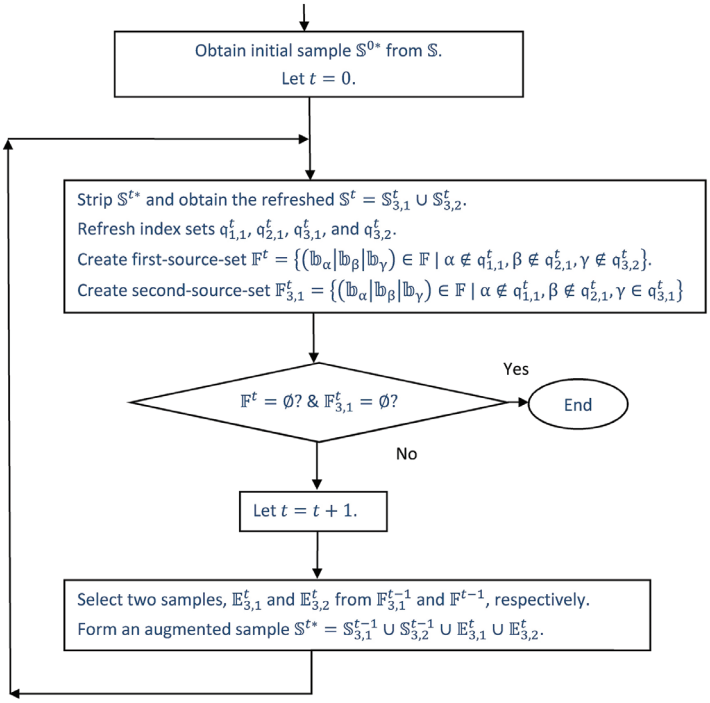

Figure 2 The principles of multiphase sampling in selecting testlet samples.

different complicated methods. For example, linear models (van der Linden, 2005) can be used to solve the problem based on TIF; the heuristic branch-and-bound method can be used to perform an intelligent search and find the solutions to the design based on TCC (Belov & Armstrong, 2008; Hansen, 1992). However, for our proposed assembly method based on testlets, if the TIF or TCC is available, the problem can be solved in a relatively more straightforward way. Unlike assembling with items or blocks, the testlets can simply be selected from a finite set of sorted testlets that meet the conditions imposed by a target TIF or TCC. This is similar to obtaining the solution to Equation 2.

Process of Stripping the Replica Blocks

Based on the previously mentioned constraints, each $\mathbb{B}1$ or $\mathbb{B}2$ block is allowed to appear in all testlets only once and each $\mathbb{B}3$ block a maximum of twice. Because the selected testlets in the combined $\mathbb{S}^{0*}$ can violate the imposed constraints, the replica blocks — either in $\mathbb{B}1$ or $\mathbb{B}2$ or the blocks appearing more than two times in $\mathbb{B}3$ — need to be stripped. The stripping process consists of the following three steps. First, strip the testlets in $\mathbb{S}^{0*}$ with $\mathbb{B}3$ blocks such that only two testlets in each block are retained in the sample. For the database used in this study, no more than 96 testlets will be retained in the sample after this stripping step. Second, the testlets with extra $\mathbb{B}2$ blocks are stripped so that only one testlet in each $\mathbb{B}2$ block is kept in the sample. After the second stripping step, fewer testlets are retained. Third, similarly, the testlets with extra $\mathbb{B}1$ blocks are stripped, and only one testlet in each $\mathbb{B}1$ block is retained. All the stripped testlets will be put back into the database. We obtain the stripped initial sample $\mathbb{S}^0$ by applying the stripping process, and in the next stage, we augment the sample $\mathbb{S}^0$. Note that in the stripping process, the order of Step 2 and Step 3 can usually be switched without noticeable effects on the final results. For the database used in this study, the sizes of the initial samples were 29 and 31 testlets, respectively, for the two examples presented in the "Results" section.

Algorithm of Multiphase Sampling

This section covers the algorithm of multiphase sampling used to augment the sample of testlets in the study. Figure 2 presents a flowchart of the strategy of sample augmentation. Several related topics are also discussed, such as termination of the algorithm, the dynamics in sample size, and the effectiveness of the algorithm.

The First Phase of Sample Augmentation

After the creation of the set of sampled forms, $\mathbb{S}^0$, the index sets need to be updated. Let $\mathfrak{q}^0_{3,1} = \left\{\gamma | \left(\mathbb{b}_\alpha \middle| \mathbb{b}_\beta \middle| \mathbb{b}_\gamma\right) \in \mathbb{S}^0 \text{ and } \mathbb{p}_\gamma \text{ appears once in } \mathbb{S}^0\right\}$ and $\mathfrak{q}^0_{3,2} = \left\{\gamma | \left(\mathbb{b}_\alpha \middle| \mathbb{b}_\beta \middle| \mathbb{b}_\gamma\right) \in \mathbb{S}^0 \text{ and } \mathbb{p}_\gamma \text{ appears twice in } \mathbb{S}^0\right\}$. Similarly, define $\mathfrak{q}^0_{1,1}$ and $\mathfrak{q}^0_{2,1}$. Define $\mathbb{S}^0_{3,1} = \left\{\left(\mathbb{b}_\alpha \middle| \mathbb{b}_\beta \middle| \mathbb{b}_\gamma\right) \in \mathbb{S}^0 \mid \gamma \in \mathfrak{q}^0_{3,1}\right\}$ and $\mathbb{S}^0_{3,2} = \left\{\left(\mathbb{b}_\alpha \middle| \mathbb{b}_\beta \middle| \mathbb{b}_\gamma\right) \in \mathbb{S}^0 \mid \gamma \in \mathfrak{q}^0_{3,2}\right\}$; $\mathbb{S}^0 = \mathbb{S}^0_{3,1} \cup \mathbb{S}^0_{3,2}$.

After the creation of the initial index sets of the block status, we have information as to whether these blocks are used. Based on $\mathfrak{q}^0_{1,1}$, $\mathfrak{q}^0_{2,1}$, $\mathfrak{q}^0_{3,1}$, and $\mathfrak{q}^0_{3,2}$, we can create a first-source set of testlets:

$$\mathbb{F}^0 = \left\{\left(\mathbb{b}_\alpha \middle| \mathbb{b}_\beta \middle| \mathbb{b}_\gamma\right) \in \mathbb{F} \mid \alpha \notin \mathfrak{q}^0_{1,1}, \beta \notin \mathfrak{q}^0_{2,1}, \gamma \notin \mathfrak{q}^0_{3,2}\right\}. \tag{3}$$

The creation of source set $\mathbb{F}^0$ and $\mathbb{F}^t$ ($t = 1, \ldots, T$) in the later phases is critical to the algorithm of sample augmentation. The set $\mathbb{F}^0$ contains the testlets with the blocks $\mathbb{b}_\gamma$ in $\mathbb{B}3$ that have never been used and is usually not empty. The augmentation process continues until $\mathbb{F}^t$ is empty. Because $\mathbb{F}^0$ is the set of testlets formed by excluding testlets with the blocks in $\mathfrak{q}^0_{1,1}$, $\mathfrak{q}^0_{2,1}$, and $\mathfrak{q}^0_{3,2}$, if a testlet $\left(\mathbb{b}_{\alpha_k} \middle| \mathbb{b}_{\beta_k} \middle| \mathbb{b}_{\gamma_k}\right)$ is drawn from $\mathbb{F}^0$, it must not be in $\mathbb{S}^0$ (i.e., $\left(\mathbb{b}_{\alpha_k} \middle| \mathbb{b}_{\beta_k} \middle| \mathbb{b}_{\gamma_k}\right) \notin \mathbb{S}^0$). Moreover, the combined set, $\mathbb{S}^0 \cup \left(\mathbb{b}_{\alpha_k} \middle| \mathbb{b}_{\beta_k} \middle| \mathbb{b}_{\gamma_k}\right)$, complies with the constraints imposed.

A second-source set of testlets can also be defined:

$$\mathbb{F}^0_{3,1} = \left\{\left(\mathbb{b}_\alpha \middle| \mathbb{b}_\beta \middle| \mathbb{b}_\gamma\right) \in \mathbb{F} \mid \alpha \notin \mathfrak{q}^0_{1,1}, \beta \notin \mathfrak{q}^0_{2,1}, \gamma \in \mathfrak{q}^0_{3,1}\right\}. \tag{4}$$

This contains the testlets with the blocks $\mathbb{p}_\gamma$ in $\mathbb{B}3$ that have been used only once. Note that $\mathbb{F}^0_{3,1}$ is usually not empty initially. Then two samples are randomly selected from $\mathbb{F}^0_{3,1}$ and $\mathbb{F}^0$:

$$\mathbb{E}^t_{3,1} = \left\{\left(\mathbb{b}_\alpha \middle| \mathbb{b}_\beta \middle| \mathbb{b}_\gamma\right) \in \mathbb{F}^0_{3,1} | \mathbb{b}_\gamma \text{ appears once}\right\} \tag{5}$$

and

$$\mathbb{E}^t_{3,2} = \left\{\left(\mathbb{b}_\alpha \middle| \mathbb{b}_\beta \middle| \mathbb{b}_\gamma\right) \in \mathbb{F}^0 | \mathbb{b}_\gamma \text{ appears up to two times}\right\}, \tag{6}$$

with constraints under optimal tactics. The sample $\mathbb{E}^1_{3,1}$, randomly drawn from $\mathbb{F}^0_{3,1}$ with constraints, consists of testlets with each block $\mathbb{b}_\gamma$ in $\mathbb{F}^0_{3,1}$ that appears only in one testlet; the sample $\mathbb{E}^1_{3,2}$, randomly drawn from $\mathbb{F}^0$ with constraints, consists of testlets with each block $\mathbb{b}_\gamma$ in $\mathbb{F}^0$ that appears two times if block $\mathbb{b}_\gamma$ appears in two or more testlets.

At the end of the first phase, we obtain an augmented sample set

$$\mathbb{S}^{1*} = \mathbb{S}^0_{3,1} \cup \mathbb{S}^0_{3,2} \cup \mathbb{E}^1_{3,1} \cup \mathbb{E}^1_{3,2}.$$

There can be replicated blocks in $\mathbb{S}^{1*}$ that violate the imposed constraints. Thus we can apply the stripping process to $\mathbb{S}^{1*}$ to obtain a stripped augmented sample $\mathbb{S}^1$. The size of $\mathbb{S}^0$ is smaller than the size of $\mathbb{S}^1$.

The tth Phase of Sample Augmentation

We continue the process of augmentation as illustrated in Figure 2 by repeating the first phase described above. Let the current phase be the tth augmentation. The previous phase is then the $(t-1)$th. The sample from the previous phase can be partitioned into two parts: $\mathbb{S}^{t-1} = \mathbb{S}^{t-1}_{3,1} \cup \mathbb{S}^{t-1}_{3,2}$, where $\mathbb{S}^{t-1}_{3,1} = \left\{\left(\mathbb{b}_\alpha \middle| \mathbb{b}_\beta \middle| \mathbb{b}_\gamma\right) \in \mathbb{S}^{t-1} \mid \gamma \in \mathfrak{q}^{t-1}_{3,1}\right\}$, and $\mathbb{S}^{t-1}_{3,2} = \left\{\left(\mathbb{b}_\alpha \middle| \mathbb{b}_\beta \middle| \mathbb{b}_\gamma\right) \in \mathbb{S}^{t-1} \mid \gamma \in \mathfrak{q}^{t-1}_{3,2}\right\}$. Based on $\mathbb{S}^{t-1}$, update $\mathfrak{q}^{t-1}_{3,1} = \left\{\gamma | \left(\mathbb{b}_\alpha \middle| \mathbb{b}_\beta \middle| \mathbb{b}_\gamma\right) \in \mathbb{S}^{t-1} \text{ and } \mathbb{b}_\gamma \text{ appears once in } \mathbb{S}^{t-1}\right\}$ and $\mathfrak{q}^{t-1}_{3,2} = \{\gamma \mid \left(\mathbb{b}_\alpha \middle| \mathbb{b}_\beta \middle| \mathbb{b}_\gamma\right) \in \mathbb{S}^{t-1} \text{ and } \mathbb{b}_\gamma \text{ appears twice in } \mathbb{S}^{t-1}\}$. Similarly, update $\mathfrak{q}^{t-1}_{1,1}$ and $\mathfrak{q}^{t-1}_{2,1}$.

Based on the refreshed $\mathfrak{q}^{t-1}_{1,1}$, $\mathfrak{q}^{t-1}_{2,1}$, and $\mathfrak{q}^{t-1}_{3,2}$, let $\mathbb{F}^{t-1} = \left\{\left(\mathbb{b}_\alpha \middle| \mathbb{b}_\beta \middle| \mathbb{b}_\gamma\right) \in \mathbb{F} \mid \alpha \notin \mathfrak{q}^{t-1}_{1,1}, \beta \notin \mathfrak{q}^{t-1}_{2,1}, \gamma \notin \mathfrak{q}^{t-1}_{3,2}\right\}$ be the first-source set. Similarly, based on the refreshed $\mathfrak{q}^{t-1}_{1,1}$, $\mathfrak{q}^{t-1}_{2,1}$, and $\mathfrak{q}^{t-1}_{3,1}$, let $\mathbb{F}^{t-1}_{3,1} = \left\{\left(\mathbb{b}_\alpha \middle| \mathbb{b}_\beta \middle| \mathbb{b}_\gamma\right) \in \mathbb{F} \mid \alpha \notin \mathfrak{q}^{t-1}_{1,1}, \beta \notin \mathfrak{q}^{t-1}_{2,1}, \gamma \in \mathfrak{q}^{t-1}_{3,1}\right\}$ be the second-source set. The process of augmentation ends if both $\mathbb{F}^{t-1}$ and $\mathbb{F}^{t-1}_{3,1}$ are empty. Otherwise, continue the augmentation and randomly select two samples from $\mathbb{F}^{t-1}_{3,1}$ and $\mathbb{F}^{t-1}$:

$$\mathbb{E}^t_{3,1} = \left\{\left(\mathbb{b}_\alpha \middle| \mathbb{b}_\beta \middle| \mathbb{b}_\gamma\right) \in \mathbb{F}^{t-1}_{3,1} | \mathbb{b}_\gamma \text{ appears once}\right\} \tag{7}$$

and

$$\mathbb{E}_{3,2}^{t} = \left\{ \left(\mathbb{b}_{\alpha} \middle| \mathbb{b}_{\beta} \middle| \mathbb{b}_{\gamma} \right) \in \mathbb{F}^{t-1} | \mathbb{b}_{\gamma} \text{ appears up to two times} \right\}, \tag{8}$$

with constraints under optimal conditions. The sample $\mathbb{E}_{3,1}^{t}$, drawn from $\mathbb{F}_{3,1}^{t-1}$ randomly with constraints under optimal conditions, as in the "Creation of an Initial Sample" section, consists of the testlets with each block $\mathbb{b}_{\gamma}$ in $\mathbb{F}_{3,1}^{t-1}$ that appears only once; the sample $\mathbb{E}_{3,2}^{t}$, drawn from $\mathbb{F}^{t-1}$ randomly with constraints under optimal conditions, as in the "Creation of an Initial Sample" section, consists of the testlets with each block $\mathbb{b}_{\gamma}$ in $\mathbb{F}^{t-1}$ that appears twice if block $\mathbb{b}_{\gamma}$ appears in two or more testlets. In each phase, the optimal strategy of the minimum deviation or the strategy in Equation 2 can be employed in selecting samples from $\mathbb{F}^{t-1}$ and $\mathbb{F}_{3,1}^{t-1}$.

At the end of the tth phase, we obtain an augmented sample set

$$\mathbb{S}^{t*} = \mathbb{S}_{3,1}^{t-1} \cup \mathbb{S}_{3,2}^{t-1} \cup \mathbb{E}_{3,1}^{t} \cup \mathbb{E}_{3,2}^{t}.$$

The stripped augmented sample $\mathbb{S}^{t}$ can be obtained by applying the stripping process to $\mathbb{S}^{t*}$. Then, start the next phase of sampling until both the first- and second-source sets are empty: $\mathbb{F}^{t} = \emptyset$ and $\mathbb{F}_{3,1}^{t} = \emptyset$.

Properties of the Algorithm

Existence of a Termination

Because total correctness requires that an algorithm terminate, the termination proof is an essential part in formal verification of the algorithm (Dijkstra & Scholten, 1980; Knuth, 1973). It is also essential to show the existence of a termination of the multiphase sampling algorithm. We stated earlier that the augmentation process ends when both the first- and second-source sets are empty: $\mathbb{F}^{t} = \emptyset$ and $\mathbb{F}_{3,1}^{t} = \emptyset$. The basic issue is whether the augmentation algorithm always leads to an end. It is straightforward to depict the algorithm when only adding one testlet at each phase, although, in practice, we always select several testlets into the sample at each phase. Let $\mathbb{S}^{t-1}$ be the stripped sample and its sample size be n^{t-1} at the $(t-1)$th phase of selection. Let $\mathbb{F}^{t-1}$ be the source set obtained. The set $\mathbb{F}^{t-1}$ is formed by excluding the testlets with the blocks in $\mathfrak{q}_{1,1}^{t-1}$, $\mathfrak{q}_{2,1}^{t-1}$, and $\mathfrak{q}_{3,2}^{t-1}$ from $\mathbb{F}$.

Assume $\mathbb{f}_{\alpha_u\beta_v\gamma_w} = \left(\mathbb{b}_{\alpha_u} \middle| \mathbb{b}_{\beta_v} \middle| \mathbb{b}_{\gamma_w} \right)$ is the only testlet drawn from $\mathbb{F}^{t-1}$ at the tth phase. Clearly $\mathbb{f}_{\alpha_u\beta_v\gamma_w} \notin \mathbb{S}^{t-1}$, so $\mathbb{S}^{t*} = \mathbb{S}^{t-1} \cup \left\{ \mathbb{f}_{\alpha_u\beta_v\gamma_w} \right\}$. Because the testlets in $\mathbb{S}^{t*}$comply with the constraints imposed, $\mathbb{S}^{t*}$ has no testlets that need to be stripped (i.e., $\mathbb{S}^{t} = \mathbb{S}^{t*}$). The sample size of $\mathbb{S}^{t}$ is $n^{t-1} + 1$. If we select one testlet at each phase, the sample $\mathbb{S}^{t}$ will be augmented with one extra testlet continuously, and the sample size will add one at each phase (i.e., for $t = 1, \ldots, T$):

$$\mathbb{S}^{1} \subset \mathbb{S}^{2} \subset \ldots \mathbb{S}^{t} \subset \ldots \mathbb{S}^{T}.$$

Simultaneously, the size of the source set will be reduced:

$$\mathbb{F}^{1} \supset \mathbb{F}^{2} \supset \ldots \mathbb{F}^{t} \supset \ldots \mathbb{F}^{T}.$$

Because the maximum sample size is 96, there exist T ($\leq$96) such that $\mathbb{F}^{T} = \emptyset$. The next stage is to select testlets from $\mathbb{F}_{3,1}^{T+1}$ ($T = 1, \ldots, T'$). We apply the same strategy to select one testlet each time from $\mathbb{F}_{3,1}^{T+1}$. Then, we update $\mathbb{F}_{3,1}^{T+1}$ and obtain $\mathbb{F}_{3,1}^{T+2}$ until $\mathbb{F}_{3,1}^{T+T'} = \emptyset$ at phase $T^{*} = T + T'$. Therefore the augmentation algorithm will terminate at T^{*}.

Dynamics in Sample Size

A case can be used to illustrate the existence of dynamics in sample size. Assume $\mathbb{f}_{\alpha_a\beta_b\gamma_c} = \left(\mathbb{b}_{\alpha_a} \middle| \mathbb{b}_{\beta_b} \middle| \mathbb{b}_{\gamma_c} \right)$ is the only testlet that contains block $\mathbb{b}_{\gamma_c}$; there are several testlets that contain block $\mathbb{b}_{\alpha_{a0}}$. If the testlet $\mathbb{f}_{\alpha_{a0}\beta_b\gamma_{c*}}$ is in the sample, $\mathbb{f}_{\alpha_{a0}\beta_b\gamma_c}$ will not be in the sample. Because the block $\mathbb{b}_{\alpha_{a0}}$ is used, the block $\mathbb{b}_{\gamma_c}$ can never appear in any testlets in the sample when the selection terminates. Otherwise, if the testlet $\mathbb{f}_{\alpha_{a0}\beta_b\gamma_c}$ is in the sample, $\mathbb{f}_{\alpha_{a*}\beta_b\gamma_{c*}}$ can also be in sample. Thus both the blocks $\mathbb{b}_{\gamma_c}$ and $\mathbb{b}_{\gamma_{c*}}$ can be used.

In general, let $\mathbb{F}^{t-1}$ be the source set obtained at the $(t-1)$th phase of selection. Let $\mathbb{f}_{\alpha_a\beta.\gamma.}$ be a selected testlet. Thus, $\mathbb{f}_{\alpha_a\beta.\gamma.}$ contains $\mathbb{b}_{\alpha_a}$ but not $\mathbb{b}_{\gamma_c}$ (i.e., $\alpha_a \in \mathfrak{q}_{1,1}$). By its definition, the source set $\mathbb{F}^{t-1}$ does not contain a specific testlet $\mathbb{f}_{\alpha_a\beta_b\gamma_c}$. Hence, the testlet $\mathbb{f}_{\alpha_a\beta_b\gamma_c}$ has no chance of being selected at the tth phase of selection. If the testlet $\mathbb{f}_{\alpha_a\beta.\gamma.}$ is never

to be stripped from the sample in a phase before the end of selection, the testlet $\mathbb{f}_{\alpha_a\beta_b\gamma_c}$ has no chance of being selected. Nevertheless, if $\mathbb{f}_{\alpha_a\beta_b\gamma_\cdot}$ is stripped from the sample in a phase before the end, the testlet $\mathbb{f}_{\alpha_a\beta_b\gamma_c}$ can be included in the source set and has a chance to be selected. In this case, because there exists uncertainty for $\mathbb{f}_{\alpha_a\beta_b\gamma_c}$ to be included in the sample, the sample size is dynamic. In this study, for example, the size of the first sample drawn is 80, and the size of the second one is 84.

Effectiveness of the Algorithm

The effectiveness of the algorithm suggests that, given a sample selected, no further testlets can be drawn from the remaining database. From the discussion, we know that the size of a final sample $\mathbb{S}^T$ is not always the maximized one among all possible samples. Nevertheless, the algorithm of multiphase sampling achieves maximum effectiveness in its searching because no testlets that can be included in the final sample $\mathbb{S}^T$ remain in the database. If a testlet $\left(\mathbb{b}_\alpha \middle| \mathbb{b}_\beta \middle| \mathbb{b}_\gamma\right)$ in the remaining database can be included, this implies two situations: First, for block $\mathbb{b}_\alpha$, $\alpha \notin \mathfrak{q}^t_{1,1}$; for $\mathbb{b}_\beta$, $\beta \notin \mathfrak{q}^t_{2,1}$; and for $\mathbb{b}_\gamma$, $\gamma \notin \mathfrak{q}^t_{3,2}$. So $\left(\mathbb{b}_\alpha \middle| \mathbb{b}_\beta \middle| \mathbb{b}_\gamma\right) \in \mathbb{F}^T$ (i.e., $\mathbb{F}^T \neq \emptyset$), which contradicts the condition of termination: $\mathbb{F}^T = \emptyset$. Second, for block $\mathbb{b}_\alpha$, $\alpha \notin \mathfrak{q}^t_{1,1}$; for $\mathbb{b}_\beta$, $\beta \notin \mathfrak{q}^t_{2,1}$; and for $\mathbb{b}_\gamma$, $\gamma \in \mathfrak{q}^t_{3,1}$. So $\left(\mathbb{b}_\alpha \middle| \mathbb{b}_\beta \middle| \mathbb{b}_\gamma\right) \in \mathbb{F}^t_{3,1}$ (i.e., $\mathbb{F}^t_{3,1} \neq \emptyset$), which contradicts the condition of termination: $\mathbb{F}^t_{3,1} = \emptyset$.

Results

In this example, two samples of testlets were drawn from the database following the procedures described in the "Database of Block-Based Testlets" section and the "Method" section. This section provides the results yielded by the multiphase sampling.

The selection of the first sample comprised two steps. The first one was creating an initial sample that was merged from two subsamples $\mathbb{S}^{0*} = \mathbb{S}^*_1 \cup \mathbb{S}^*_2$: a set of 137 testlets yielded by selecting one testlet for each $\mathbb{B}1$ block and a set of 96 testlets yielded by selecting up to two testlets for each $\mathbb{B}3$ block. In addition, to improve the psychometric properties of the sampled testlets, these two subsamples were drawn based on the optimal tactic of minimum deviation in selecting $\mathbb{S}^*_1$ and $\mathbb{S}^*_2$. The deviation in $\mathbb{B}1$ is defined as $d_{\mathbb{b}_\alpha} = a_{\mathbb{b}_\alpha} - \overline{a}_{\mathbb{b}_{\alpha\cdot}}$, where $a_{\mathbb{b}_\alpha}$ is the mean-*a* parameter of each block and $\overline{a}_{\mathbb{b}_{\alpha\cdot}}$ is the overall average across all blocks. For each block $\mathbb{b}_{\alpha_u}$ in $\mathbb{B}1$, from the subset of the testlets in the source sets that contained $\mathbb{b}_{\alpha_u}$, we selected a testlet with the minimum deviation. On the basis of the same optimal principle, we selected up to two testlets for each $\mathbb{B}3$ block. The averages of the mean-*a* parameters of the blocks in $\mathbb{B}1$, $\mathbb{B}2$, and $\mathbb{B}3$ can be found in Tables 1 – 3. Then, according the constraints imposed, the stripping process was applied to drop the replica blocks in $\mathbb{B}1$ and $\mathbb{B}2$; the size of the stripped initial sample $\mathbb{S}^0$ was 29.

The second step involved using the algorithm of multiphase sampling to augment the stripped initial sample. At each phase, say, *t*th, two samples, $\mathbb{E}^t_{3,1}$ and $\mathbb{E}^t_{3,2}$, were sampled from subsource sets $\mathbb{F}^{t-1}_{3,1}$ and $\mathbb{F}^{t-1}$ with constraints. On the basis of the processing records of the Phase 1 selection, the sizes of the source sets $\mathbb{F}^{t-1}_{3,1}$ and $\mathbb{F}^{t-1}$ were 6,248 and 2,810, and the sample sizes of $\mathbb{E}^1_{3,1}$ and $\mathbb{E}^1_{3,2}$ were 19 and 34, respectively. Although the strategy in Equation 2 was not difficult to implement, the optimal principle of the minimum deviation was also applied in each selection phase in this step because the samples presented in this section were used to demonstrate that the algorithm could accomplish the goal of sampling testlets with constraints. After four phases of sample augmentation, the size of the final sample ($\mathbb{S}^4$) of testlets obtained was 80. Across all 80 testlets, the $\mathbb{B}1$ and $\mathbb{B}2$blocks appeared only once, and the $\mathbb{B}3$ blocks appeared up to two times.

The selection of the second sample was based on the database that excluded all of the testlets in the first sample. The same sampling procedure was used to select the sample. The size of the stripped initial sample was 31. After four phases of sample augmentation, the final size of the second sample was 84. The selection in each phase also applied the optimal strategy of minimum deviation.

The difference in the sizes between two empirical samples in this case showed the dynamics in sample size, as shown in the "Properties of the Algorithm" section. The algorithm can be improved in a few ways. For example, when several samples need to be drawn, the algorithm can be refined to draw a sample with larger sample size in an earlier selection phase than in a later one. Nevertheless, this strategy may not be the goal of a test operation.

Summary

In this study, the algorithm of multiphase sampling was developed to select block-based testlets (or test forms). In the sampling process, different types of constraints were imposed on the blocks of the testlets in the selected samples. The first step was to form an initial sample that consists of two subsamples drawn with controlled selection (Cochran, 1977, p. 124) of blocks' appearance in forms. An optimal strategy was employed to achieve balance in the IRT parameters for the items in blocks. The second step was to apply the multiphase sampling to augmenting the initial sample and to yield a sample of testlets as large as possible. In each phase of sample augmentation, a source set of testlets was created containing those testlets made up of blocks that had not yet been used. Next, the testlets with replicated blocks were stripped. Then, the information of the selection status of all the blocks was updated. Several topics related to the algorithm were also discussed, including the termination of the algorithm, the dynamics in sample size, and the effectiveness of the algorithm.

The sampling process enables us to simultaneously draw as large a sample of testlets as possible with constraints enforced, whereas other computational methods usually yield one or several test forms at each run, yet with no possibility of attaining the maximum batch size with imposed constraints, as the proposed algorithm of multiphase sampling does. There are no problems of overusing or underusing blocks for the selected testlets. Although the development of the process was based on the constraints described in the "Constraints Imposed on Testlets" section, the data in this study were used for illustrative purposes, and the algorithm can be readily adapted to altered constraints. Without loss of generality, three alternatives are considered here:

- Case 1: Set the appearance of each $\mathbb{B}3$ block in the sampled testlets up to three times, with all other constraints remaining the same. At the tth phase of sample augmentation, we need to add $\mathbf{q}_{3,3}^{t-1}$ to $\mathbf{q}_{3,1}^{t-1}$ and $\mathbf{q}_{3,2}^{t-1}$; then $\mathbb{S}^{t-1} = \mathbb{S}_{3,1}^{t-1} \cup \mathbb{S}_{3,2}^{t-1} \cup \mathbb{S}_{3,3}^{t-1}$. Next, define source sets $\mathbb{F}^{t-1}$, $\mathbb{F}_{3,1}^{t-1}$, and $\mathbb{F}_{3,2}^{t-1}$ and draw three samples $\mathbb{E}_{3,1}^{t}$, $\mathbb{E}_{3,2}^{t}$, and $\mathbb{E}_{3,3}^{t}$ from $\mathbb{F}_{3,1}^{t-1}$, $\mathbb{F}_{3,2}^{t-1}$, and $\mathbb{F}^{t-1}$, respectively. At the end of the tth phase, the stripped augmented sample $\mathbb{S}^{t}$ can be obtained by applying the stripping process to $\mathbb{S}^{t*} = \mathbb{S}_{3,1}^{t-1} \cup \mathbb{S}_{3,2}^{t-1} \cup \mathbb{S}_{3,3}^{t-1} \cup \mathbb{E}_{3,1}^{t} \cup \mathbb{E}_{3,2}^{t} \cup \mathbb{E}_{3,3}^{t}$. For more details, see the appendix.
- Case 2: Set the appearance of each $\mathbb{B}3$ block in the testlets sampled to be only once, instead of up to twice, with all other constraints remaining the same. The practice is also straightforward. At the tth phase of sample augmentation, we only create $\mathbf{q}_{3}^{t-1}$, instead of $\mathbf{q}_{3,1}^{t-1}$ and $\mathbf{q}_{3,2}^{t-1}$; then $\mathbb{S}^{t-1} = \mathbb{S}_{3}^{t-1}$. After the source set $\mathbb{F}_{3}^{t-1}$ is created, the sample $\mathbb{E}_{3}^{t}$ will be selected from $\mathbb{F}_{3}^{t-1}$.
- Case 3: Set each block in all positions (either in $\mathbb{B}1$, $\mathbb{B}2$, or $\mathbb{B}3$) in the sampled testlets to appear up to two times. We can select just two samples of testlets with each block to appear only once in all three positions—in Case 2—and then merge two samples.

In survey sampling, multiphase sampling is usually applied to obtaining information about stratification and estimation. The algorithm of multiphase sampling developed in this study was designed to sample testlets with constraints, and multiphase sampling was applied to augmenting the sample and updating the selection status of blocks, repeatedly, until the process ended. Because of the effectiveness of the algorithm (see the "Properties of the Algorithm" section), the process was optimized: No testlets that remained in the database had nonselected blocks when the algorithm terminated. Therefore high efficiency was ascertained for the algorithm. Other properties of the algorithm were also discussed in this report, including the termination of the algorithm and the dynamics in sample size. For future research, the algorithm can be further improved in a few ways, according to the different assessment conditions, as discussed in the "Results" section.

Acknowledgments

The authors thank James Carlson, Daniel McCaffrey, Shelby Haberman, Usama Ali, and Peter van Rijn for their suggestions and comments. The authors also thank Kim Fryer and Ayleen Gontz for editorial help. Any opinions expressed in this paper are those of the authors and not necessarily those of Educational Testing Service.

References

Ali, U. S., & van Rijn, P. W. (2016). An evaluation of different statistical targets for assembling parallel forms in item response theory. *Applied Psychological Measurement*, *40*, 163–179. https://doi.org/10.1177/0146621615613308

Bellman, R. (1957). *A Markovian decision process* (Technical report). DTIC Document. Santa Monica, CA: the RAND Corporation.

Belov, D. I., & Armstrong, R. D. (2008). A Monte Carlo approach to the design, assembly and evaluation of multi-stage adaptive tests. *Applied Psychological Measurement*, *32*, 119 – 137. https://doi.org/10.1177/0146621606297308

Bennett, R. E. (1998). *Reinventing assessment: Speculations on the future of large-scale educational testing* (ETS Policy Information Center Report). Princeton, NJ: Educational Testing Service.

Chang, H. H., Qian, J., & Ying, Z. (2001). Alpha-stratified multistage computerized adaptive testing with beta blocking. *Applied Psychological Measurement*, *25,* 333 – 341. https://doi.org/10.1177/01466210122032181

Chang, H. H., & Zhang, J. (2002). Hypergeometric family and item overlap rates in computerized adaptive testing. *Psychometrika*, *67*, 387 – 398. https://doi.org/10.1007/BF02294991

Chen, P. H., Chang, H. H., & Wu, H. (2012). Item selection for the development of parallel forms from an IRT-based seed test using a sampling and classification approach. *Educational and Psychological Measurement*, *72*, 933 – 953. https://doi.org/10.1177/0013164412443688

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York, NY: John Wiley.

Davis, L. L., & Dodd, B. G. (2003). Item exposure constraints for testlets in the verbal reasoning section of the MCAT. *Applied Psychological Measurement*, *27*, 335 – 356. https://doi.org/10.1177/0146621603256804

Debeer, D., Ali, U., & van Rijn, P. W. (2017). Evaluating statistical targets for assembling parallel mixed-format test forms. *Journal of Educational Measurement*, *54*, 218 – 242. https://doi.org/10.1111/jedm.12142

Dijkstra, E. W., & Scholten, C. S. (1980). Termination detection for diffusing computations. *Information Processing Letters*, *11*(1), 1 – 4. https://doi.org/10.1016/0020-0190(80)90021-6

Foster, D. F., & Miller, H. L., Jr. (2012). Global test security issues and ethical challenges. In A. Ferrero, Y. Korkut, M. M. Leach, G. Lindsay, & M. J. Stevens (Eds.), *The Oxford handbook of international psychological ethics* (pp. 216 – 232). Oxford, England: Oxford University Press.

Hansen, E. R. (1992). *Global optimization using interval analysis*. New York, NY: Dekker.

Kish, L. (1965). *Survey sampling*. New York, NY: John Wiley.

Knuth, D. E. (1973). *The art of computer programming: Vol. 1. Fundamental algorithms* (2nd ed.). Reading, MA: Addison-Wesley.

Leung, C. K., Chang, H. H., & Hau, K. T. (2002). Item selection in computerized adaptive testing: Improving the *a*-stratified design with the Sympson – Hetter algorithm. *Applied Psychological Measurement*, *26*, 376 – 392. https://doi.org/10.1177/014662102237795

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Luecht, R. M. (2012). Computer-based and computer-adaptive testing. In M. Simon, K. Ercikan, & M. Rousseau (Eds.), *Improving large-scale assessment in education: Theory, issues, and practice* (pp. 62 – 84). Philadelphia, PA: Taylor and Francis.

Luecht, R. M., & Hirsch, T. M. (1991, June). *Computerized test construction of parallel forms with problem-linked items*. Paper presented at the meeting of the Psychometric Society, New Brunswick, NJ.

Luecht, R. M., & Hirsch, T. M. (1992). Computerized test construction using average growth approximation of target information functions. *Applied Psychological Measurement*, *16*, 41 – 52. https://doi.org/10.1177/014662169201600104

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 257 – 306). Westport, CT: Praeger.

Rittaud, B., & Heeffer, A. (2014). The pigeonhole principle, two centuries before Dirichlet. *Mathematical Intelligencer*, *36*, 27 – 29. https://doi.org/10.1007/s00283-013-9389-1

Stocking, M. L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, *23*, 57 – 75. https://doi.org/10.3102/10769986023001057

Swanson, L., & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement*, *17*, 151 – 166. https://doi.org/10.1177/014662169301700205

van der Linden, W. J. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement*, *22*, 195 – 211. https://doi.org/10.1177/01466216980223001

van der Linden, W. J. (2003). Some alternatives to Sympson – Hetter item-exposure control in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, *28*, 249 – 265. https://doi.org/10.3102/10769986028003249

van der Linden, W. J. (2005). *Linear models for optimal test design*. New York, NY: Springer. https://doi.org/10.1007/0-387-29054-0

van der Linden, W. J., & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics*, *29*, 273 – 291. https://doi.org/10.3102/10769986029003273

Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, *67*, 575 – 588. https://doi.org/10.1007/BF02295132

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*, 185 – 201. https://doi.org/10.1111/j.1745-3984.1987.tb00274.x

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement*, *27*, 1 – 14. https://doi.org/10.1111/j.1745-3984.1990.tb00730.x

Wollack, J. A., & Fremer, J. J. (2013). *Handbook of test security*. New York, NY: Routledge. https://doi.org/10.4324/9780203664803

Appendix

Sampling With Constraint on Testlets With Each $\mathbb{B}3$ Block's Appearance up to Three Times

The altered constraint sets the appearance of each $\mathbb{B}3$ block in the sampled testlets up to three times with all other constraints remaining the same, as in the section "Index Set of Sampled Blocks." The creation of the initial sample is similar to those in the section "Creating an Initial Sample," except that the subsample $\mathbb{S}_1^*$ has up to three testlets for each $\mathbb{B}3$ block. Similar changes need to be made for the stripping process of the replica of blocks.

At the tth phase of sample augmentation, we need to add an index set $\mathfrak{q}_{3,3}^{t-1}$ in addition to $\mathfrak{q}_{3,1}^{t-1}$ and $\mathfrak{q}_{3,2}^{t-1}$. The sample from the previous phase will be partitioned into three parts: $\mathbb{S}^{t-1} = \mathbb{S}_{3,1}^{t-1} \cup \mathbb{S}_{3,2}^{t-1} \cup \mathbb{S}_{3,3}^{t-1}$. Define the first-source set as

$$\mathbb{F}^{t-1} = \left\{ \left(\mathbb{b}_\alpha \middle| \mathbb{b}_\beta \middle| \mathbb{b}_\gamma \right) \in \mathbb{F} \mid \alpha \notin \mathfrak{q}_{1,1}^{t-1}, \beta \notin \mathfrak{q}_{2,1}^{t-1}, \gamma \notin \mathfrak{q}_{3,3}^{t-1} \right\},$$

the second-source set as

$$\mathbb{F}_{3,1}^{t-1} = \left\{ \left(\mathbb{b}_\alpha \middle| \mathbb{b}_\beta \middle| \mathbb{b}_\gamma \right) \in \mathbb{F} \mid \alpha \notin \mathfrak{q}_{1,1}^{t-1}, \beta \notin \mathfrak{q}_{2,1}^{t-1}, \gamma \in \mathfrak{q}_{3,1}^{t-1} \right\},$$

and the third-source set as

$$\mathbb{F}_{3,2}^{t-1} = \left\{ \left(\mathbb{b}_\alpha \middle| \mathbb{b}_\beta \middle| \mathbb{b}_\gamma \right) \in \mathbb{F} \mid \alpha \notin \mathfrak{q}_{1,1}^{t-1}, \beta \notin \mathfrak{q}_{2,1}^{t-1}, \gamma \in \mathfrak{q}_{3,2}^{t-1} \right\}.$$

Finally, select three samples from $\mathbb{F}_{3,1}^{t-1}$, $\mathbb{F}_{3,2}^{t-1}$, and $\mathbb{F}^{t-1}$:

$$\mathbb{E}_{3,1}^{t} = \left\{ \left(\mathbb{b}_\alpha \middle| \mathbb{b}_\beta \middle| \mathbb{b}_\gamma \right) \in \mathbb{F}_{3,1}^{t-1} | \mathbb{b}_\gamma \text{ appears once} \right\},$$

$$\mathbb{E}_{3,2}^{t} = \left\{ \left(\mathbb{b}_\alpha \middle| \mathbb{b}_\beta \middle| \mathbb{b}_\gamma \right) \in \mathbb{F}_{3,2}^{t-1} | \mathbb{b}_\gamma \text{ appears up to two times} \right\},$$

$$\mathbb{E}_{3,3}^{t} = \left\{ \left(\mathbb{b}_\alpha \middle| \mathbb{b}_\beta \middle| \mathbb{b}_\gamma \right) \in \mathbb{F}^{t-1} | \mathbb{b}_\gamma \text{ appears up to three times} \right\}.$$

At the end of the tth phase, we obtain an augmented sample set

$$\mathbb{S}^{t*} = \mathbb{S}_{3,1}^{t-1} \cup \mathbb{S}_{3,2}^{t-1} \cup \mathbb{S}_{3,3}^{t-1} \cup \mathbb{E}_{3,1}^{t} \cup \mathbb{E}_{3,2}^{t} \cup \mathbb{E}_{3,3}^{t}.$$

The stripped augmented sample $\mathbb{S}^t$ can be obtained by applying the stripping process to $\mathbb{S}^{t*}$. The augmentation process will continue by applying the next phase of sampling until all three source sets are empty: $\mathbb{F}_{3,1}^{T} = \emptyset$, $\mathbb{F}_{3,2}^{T} = \emptyset$, and $\mathbb{F}^{T} = \emptyset$ at phase T.

Suggested citation:

Qian, J., Gu, L., & Li, S. (2019). *Applying multiphase sampling to selecting testlets with constraints on item blocks* (Research Report No. RR-19-03). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12239

Action Editor: James Carlson

Reviewers: Usama Ali and Peter van Rijn